HumanBase Documentation

Release 1.0

User Guide

Apr 13, 2023

Contents

CHAPTER 1

About

HumanBase is a "one stop shop" for biological and biomedical researchers interested in data-driven predictions of gene expression, function, regulation, and interactions in human, particularly in the context of specific cell types/tissues, development, and human disease.

Data-driven integrative analyses are especially powerful because they reach beyond "known biological knowledge" represented in the biological literature to identify novel associations that are not biased toward well-studied areas of biomedical research. Thus, carefully designed algorithms can drive the development of experimentally testable hypotheses, enabling deeper understanding of basic biology at the molecular level, pathophysiology, and paving the way to therapy and drug development.

CHAPTER 2

Example use case

A researcher who studies the role of the immune system and inflammation in chronic kidney disease wants to identify candidate genes for these disorders. Unfortunately, as with most specific disease contexts outside of cancer, few datasets are available for these diseases, none are focused on the role of inflammation or the immune system, and no dataset is specific to her cell-lineage of interest. Even identifying which genes are expressed in the cell type relevant to glomerular disease (podocytes) is currently impossible as this cell lineage cannot be isolated for high-throughput experiments in human.

Using HumanBase, she will be able to examine data-driven predictions of genes expressed in the podocyte cells and analyze predicted functional and mechanistic networks specific to the kidney glomerulus. She could also provide the system with a list of relevant GWAS or family-based study results and the system will reprioritize these results based on the relevant functional maps. She will be able to iteratively refine this analysis by limiting the data used in the integration only to kidney datasets or by integrating her own data in the analysis.

CHAPTER 3

Licensing

All data in HumanBase are freely available under a CC-BY 4.0 license. Please give appropriate credit, provide a link to the license, and indicate if changes were made.

CHAPTER 4

Who are we?

HumanBase is actively developed by the Genomics group at the Flatiron Institute .

CHAPTER 5

Help topics

5.1 Getting Started

5.2 Functional Networks

In order to leverage the vast collections of raw, noisy genomic data, they must be integrated, summarized, and presented in a biologically informative manner. We provide a means of mining tens of thousands of whole-genome experiments by way of functional interaction networks. Each interaction network represents a body of data, probabilistically weighted and integrated, focused on a particular biological question. These questions can include, for example, the function of a gene, the relationship between two pathways, or the processes disrupted in a genetic disorder. (Huttenhower, et. al 2008)

5.2.1 Method

Briefly, functional integration relies on the construction of process-specific functional relationship networks. These are interaction networks in which each node represents a gene, each edge a functional relationship, and an edge between two genes is probabilistically weighted based on experimental evidence relating to those genes. We integrate evidence from many data sets, with each data set weighted in a process-specific manner.

One naïve Bayesian classifier is trained per biological area of interest (e.g. a tissue, or a specific biological process), using the appropriate gold standard for the biological context in addition to one global process-unaware classifier trained using the complete gold standard. Each classifier f consisted of a class node predicting the binary presence or absence of a functional relationship (FR) between two genes and n nodes conditioned on FR, each representing the value of a data set Dk.

Parameter regularization is performed as described in Steck and Jaakkola (2002) using mutual information between data sets to estimate a strength of prior belief for each data set. While a large amount of shared information does not guarantee a redundant data set, since the same subset of information could be shared many times, it provides a valuable quantitative estimate of data set uniqueness.

5.2.2 Genomics data types

We collected and integrated 987 genome-scale data sets encompassing approximately 38,000 conditions from an estimated 14,000 publications including both expression and interaction measurements.

- Gene co-expression: All gene expression data sets are from NCBI's Gene Expression Omnibus (GEO). Genes with more than 30% of values missing were removed, and remaining missing values were imputed using ten nearest neighbors. Non-log-transformed data sets were log transformed. Expression measurements were summarized to Entrez identifiers, and duplicate identifiers were merged. The Pearson correlation was calculated for each gene pair, normalized with Fisher's z transform, mean subtracted and divided by the standard deviation.
- Protein-interaction: Interaction data are collected from BioGRID, IntAct, MINT, and MIPS.
- TF regulation: To estimate shared transcription factor regulation between genes, we collected binding motifs from JASPAR. Genes were scored for the presence of transcription factor binding sites using the MEME software suite. Motif matches were treated as binary scores (present if $P < 0.001$). The final score for each gene pair was obtained by calculating the Pearson correlation between the motif association vectors for the genes.
- MSigDB purturbations and miRNA: Chemical and genetic perturbation (c2:CGP) and microRNA target (c3:MIR) profiles were downloaded from the Molecular Signatures Database (MSigDB). Each gene pair's score was the sum of shared profiles weighted by the specificity of each profile

5.2.3 Evidence

The "evidence" for an edge is measured as the contribution or "influence" of each dataset on the posterior classification probability. Each dataset contribution is calculated as the posterior probability of a functional relationship given only that dataset, minus the prior probablility.

Contribution of dataset D to an edge functional relationship prediction (FR):

```
contribution(D) = P(FR | D) - P(FR)
```

Note that the contributions will not sum to 1.0, as each contribution is measured separately. Generally, individual gene expression datasets will not contribute much to the posterior probability but cumulatively can make a significant contribution.

5.3 Tissue-specific Networks

The precise actions of genes are frequently dependent on their tissue context, and human diseases result from the disordered interplay of tissue- and cell lineage–specific processes. These factors combine to make the understanding of tissue-specific gene functions, disease pathophysiology and gene-disease associations particularly challenging.

5.3.1 Functional interactions

HumanBase builds genome-scale functional maps of human tissues by integrating a collection of data sets covering thousands of experiments contained in more than 14,000 distinct publications. To integrate these data, we automatically assess each data set for its relevance to each of 144 tissue- and cell lineage–specific functional contexts. The resulting functional maps provide a detailed portrait of protein function and interactions in specific human tissues and cell lineages ranging from B lymphocytes to the renal glomerulus and the whole brain. This approach allows us to profile the specialized function of genes in a high-throughput manner, even in tissues and cell lineages for which no or few tissue-specific data exist.

These maps can answer biological questions that are specific to a single gene in a single tissue. For example, we have used these maps for the gene IL1B (encoding interleukin (IL)-1β) in the blood vessel network, where it has a key role in inflammation, to predict lineage-specific responses to IL-1β stimulation, which we experimentally confirmed.

5.3.2 Examples

IL1B in blood vessel

We examined and experimentally verified the tissue-specific molecular response of blood vessel cells to stimulation by IL-1β (IL1B), a pro-inflammatory cytokine. We anticipated that the genes most tightly connected to IL1B in the blood vessel network would be among those responding to IL-1β stimulation in blood vessel cells. We tested this hypothesis by profiling the gene expression of human aortic smooth muscle cells (HASMCs; the predominant cell type in blood vessels) stimulated with IL-1β.

Examination of the genes whose expression was significantly upregulated at 2 h after stimulation showed that 18 of the 20 IL1B network neighbors were among the top 500 most upregulated genes in the experiment (P = 2.07 × 1023). The blood vessel network was the most accurate tissue network in predicting this experimental outcome; none of the other 143 tissue-specific networks or the tissue-naive network performed as well when evaluated by each network's ability to predict the result of IL-1β stimulation on the cells.

Fig. 1: We anticipated that the genes most tightly connected to IL1B in the blood vessel network would be among those responding to IL-1β stimulation in blood vessel cells (a) The 20 genes most tightly connected to IL1B in the blood vessel network are shown. These genes are predicted to respond to IL-1β stimulation in blood vessel. (b) The bar plot shows the differential expression levels of the 20 IL1B neighbors measured in a microarray experiment at 0 h and 2 h after IL-1β stimulation in HASMCs, which constitute most of the blood vessel. Each bar represents the gene's log ratio of mean expression at 2 h over its mean expression at 0 h. Error bars represent regularized pooled standard errors estimated by LIMMA (n = 4 biological replicates). Eighteen of the 20 IL1B network neighbors (labeled in bold) were found to be among the most significantly differentially expressed genes at 2 h relative to 0 h (P = 1.95 × 1023).

5.4 Functional module detection

HumanBase applies community detection to find cohesive gene clusters from a provided gene list and a selected relevant tissue. Genes within a cluster share local network neighborhoods and together form a cohesive, specific functional module. Module detection enables systematic association of genes - even functionally uncharacterized genes - to specific processes and phenotypes represented in the detected modules. Functional modules are identified with tissue-specific networks, which predict gene interactions from massive data collections. Thus the discovered modules potentially capture higher-order tissue-specific function.

5.4.1 Method

The approach[1] is based on shared k-nearest-neighbors (SKNN) and the Louvain community-finding algorithm to cluster the user-selected tissue network into distinct modules of tightly connected genes. The SKNN-based strategy has the advantages of alleviating the effect of high-degree genes and accentuating local network structure by connecting genes that are likely to be functionally clustered together.

This technique proceeds as follows:

(i) First, we create a subset of the user-selected network containing only the user-provided genes and all the edges between them. Given the resulting graph G with V nodes (user-provided genes) and E edges, with each edge between genes i and j associated with a weight p_{ij},

(ii) Calculate a new weight for the edge between each pair of nodes i and j that is equal to the number of k nearest neighbors (based on the original weights p_{ij}) shared by i and j;

(iii) Choose the top 5% of the edges based on the new edge weights, and apply a graph clustering algorithm.

This approach has two key desirable characteristics:

(i) Choosing the highest k values instead of all edges deemphasizes high-degree 'hub' nodes and brings equal attention to highly specific edges between low-degree nodes;

(ii) Emphasizing local network-structure by connecting nodes that share a number of local neighbors automatically links genes that are highly likely to be part of the same cluster.

We use a dynamic `k = min(50, 0.2 * |V|)` to obtain the shared-nearest-neighbor tissue-specific network and apply the Louvain algorithm to cluster this network into distinct modules. To stabilize clustering across different runs of the Louvain algorithm, we run the algorithm 100 times and calculate cluster comembership scores for each pair of genes that was equal to the fraction of times (out of 100) the pair was assigned to the same cluster. Genes are assigned to clusters where their comembership score 0.9.

Resulting modules are then tested for functional enrichment using genes annotated to Gene Ontology biological process terms. Representative processes and pathways enriched within each cluster are presented alongside of the cluster with their resulting Q value. The Q value of each term associated to the modules is calculated using one-sided Fisher's exact tests and Benjamini–Hochberg corrections to correct for multiple tests.

1. Krishnan A*, Zhang R*, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG.(2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature Neuroscience.

5.5 NetWAS - Network-wide Association Study

Tissue-specific networks provide a new means to generate hypotheses related to the molecular basis of human disease. We developed an approach, termed network-wide association study (NetWAS). In NetWAS, the statistical associations from a standard GWAS guide the analysis of functional networks. This reprioritization method is driven by discovery and does not depend on prior disease knowledge. NetWAS, in conjunction with tissue-specific networks, effectively reprioritizes statistical associations from distinct GWAS to identify disease-associated genes, and tissue-specific NetWAS better identifies genes associated with hypertension than either GWAS or tissue-naive NetWAS.

5.5.1 Method

NetWAS trains a support vector machine classifier using nominally significant ($P < 0.01$) genes as positive examples and 10,000 randomly selected non-significant (P 0.01) genes as negatives. The classifier is constructed using a tissue network relevant to a disease (e.g. kidney for hypertension), where the features of the classifier are the edge weights of the labeled examples to all the genes in the network. Genes are re-ranked using their distance from the hyperplane, which represent a network-based prioritization of a GWAS, termed NetWAS.

To calculate per-gene P values for a GWAS, we suggest the versatile gene-based association study (VEGAS) system.

We have performed and evaluated NetWAS on six GWAS: C-reactive protein levels (lnCRP), type 2 diabetes (T2D), body mass index (BMI), hypertension (ht), alzheimer's (adni) and advanced age-related macular degeneration (advanced AMD).

5.5.2 GWAS File

NetWAS requires as input a GWAS result file, with per-gene p-values. We suggest the versatile gene-based association study (VEGAS) system for calculating gene p-values, but we also support forge and pseq formats.

- VEGAS: versatile gene-based association study
- FORGE: multivariate calculation of gene-wide p-values from Genome-Wide Association Studies Authors and Affiliations
- PLINK/SEQ: a library for the analysis of genetic variation data

  Note that the expected format is from the output of Gene/group-based association tests

5.5.3 NetWAS Results

When a NetWAS analysis finishes, a result file will be emailed to the provided address and/or can be accessed at a given URL. An example file is show below:

```
##################################################################################
# HumanBase NetWAS Analysis Results
#
# Job id:      d7732f19-916d-4458-97b5-936b8d6345cb
# Job title:
# Email:
# Created:     2017-08-21 17:07:33 EDT
# GWAS file:   bmi-2012.out.txt
# GWAS format: vegas
# Tissue:      adipose_tissue
# P-value:     0.01
#
# Result file format:
#
# Column 1) Gene symbol
# Column 2) Training label: 1 (+, nominally significant p-value)
#                          -1 (-, not nominally significant p-value)
#                           0 (not used in training)
# Column 3) NetWAS Score: Distance from the SVM separating hyperplane. Positive scores
# are in the positive direction (more like nominally significant), negative scores
# are in the negative direction (more like non-significant)
##################################################################################
# NetWAS citation:
# Greene CS*, Krishnan A*, Wong AK*, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang
# R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K,
# Grosser T, Troyanskaya OG. (2015). Understanding multicellular function and
# disease with human tissue-specific networks. Nature Genetics. 10.1038/ng.3259w.
##################################################################################
KRT6B  -1      0.561327
EMP1   -1      0.541169
ZBTB41 -1      0.503238
PNPLA8 -1      0.454396
ITGB4  -1      0.440985
........
```

5.5.4 Examples

Hypertension GWAS

Hypertension is a major cardiovascular risk factor and a complex trait involving a large number of genetic variants. We converted SNP-level association statistics into gene-level statistics for each of three recorded phenotypes—diastolic blood pressure (DBP), systolic blood pressure (SBP) and hypertension. Using the tissue-specific network for kidney, a tissue that has a central role in blood pressure control, NetWAS constructed a classifier that identified tissue-specific network connectivity patterns associated with the phenotype of interest. Genes annotated to hypertension phenotypes in the Online Mendelian Inheritance in Man (OMIM) database were more highly ranked by this classifier than by the initial GWAS. (citation)

Fig. 2: Genes ranked using GWAS (gray) and genes reprioritized using NetWAS (brown) were assessed for correspondence to genes known to be associated with hypertension phenotypes, regulatory processes and therapeutics. We compared individual (systolic blood pressure, SBP; diastolic blood pressure, DBP; hypertension, HTN) as well as combined hypertension endpoints. (a) Gene rankings were compared to OMIM-annotated hypertension genes using AUC. The AUC for the tissue-specific NetWAS is consistently higher than that for the original GWAS for all hypertension endpoints. Merging the network-based predictions for the three hypertension-related endpoints into a combined phenotype results in the best performance (AUC = 0.77; original GWAS AUC = 0.62; the dashed line at 0.5 denotes the AUC of a baseline random predictor). (b,c) Gene rankings were also assessed for enrichment of genes involved in the regulation of blood pressure (GO) (b) and targets of antihypertensive drugs (DrugBank) (c). The top NetWAS results were significantly enriched for genes involved in blood pressure regulation as well as for genes that are targets of antihypertensive drugs. Enrichment was calculated as a z score (Online Methods), with higher scores indicating a greater shift from the expected ranking toward the top of the list. In nearly all cases, the NetWAS ranking was both significantly enriched with the respective gene sets (z score > 1.645 P value < 0.05) and more enriched than in the original GWAS ranking.

Additional GWAS

Fig. 3: Each bar shows the performance of NetWAS reprioritization as measured by the area under the curve (AUC) of documented disease associations with the disease specified in the label above the plot. The horizontal axis shows relevant networks (colored bars) and GWAS alone (gray bars), and the horizontal axis label describes the GWAS phenotype from which associations were obtained.

5.6 Sei

5.6.1 Introduction

Sei is a deep-learning-based framework for systematically predicting sequence regulatory activities and applying sequence information to understand human genetics data. Sei provides a global map from any sequence to regulatory activities, as represented by 40 sequence classes. Each sequence class integrates predictions for 21,907 chromatin profiles (transcription factor, histone marks, and chromatin accessibility profiles across a wide range of cell types) from the underlying Sei deep learning model. You can also find the Sei code repository here or read about our manuscript here.

Sequence class-level variant effects are computed by comparing the predictions for the reference and the alternative alleles. A positive score indicates an increase in sequence class activity by the alternative allele and vice versa. Sequence class-level scores are computed by projecting the 21,907 chromatin profile predictions for the sequence to the unit vector that represents each sequence class.

For older DeepSEA models see: *DeepSEA (Beluga)* (2019)

5.6.2 Input

File formats

We support three types of input: vcf, fasta, bed. If you want to predict effects of noncoding variants, use vcf format input. If you want to predict chromatin feature probabilities for DNA sequences, use fasta format. If you want to specify sequences from the human reference genome (GRCh37/hg19), you can use bed format. See below for a quick introduction:

VCF format is used for specifying a genomic variant. A minimal example is `chr1 109817590 - G T` (if you want to copy cover this text as input, you will need to change spaces to tabs). The five columns are chromosome, position, name, reference allele, and alternative allele. Currently, the genome position needs to be in GRCh37/hg19

Fasta format input should include sequences of 4096bp length each. If a sequence is longer than 4096bp, only the center 4096bp will be used.

Bed format provides another way to specify sequences in human reference genome (hg19). The bed input should specify 4096bp-length regions. A minimal example is `chr1 109817091 109821186`. The three columns are chromosome, start position, and end position.

Genome coordinates

We support only `GRCh37/hg19` genome coordinates. You can use LiftOver to convert your coordinates to the correct version.

Large submissions

We recommend using the web server if you have <10,000 variants or sequences. You will experience degraded performance when submitting a larger set of sequences. In those instances, we suggest that you split the set into multiple <10,000 submissions, or run the standalone version on your local machine, or contact our group directly.

5.6.3 Output

Sequence classes

The Sei framework predicts 40 sequence class scores, covering a wide range of regulatory activities such as cell-type-specific enhancers and promoters, as well as 21,907 chromatin profiles for any DNA sequence.

To help interpretation, we grouped sequence classes into groups including P (Promoter), E (Enhancer), CTCF (CTCF-cohesin binding), TF (TF binding), PC (Polycomb-repressed), HET (Heterochromatin), TN (Transcription), and L (Low Signal) sequence classes. Please refer to our manuscript for a more detailed description of the sequence classes.

Note: sequence class predictions are only available for vcf inputs.

```
| Sequence class label |                Sequence class name | Rank by size | Group |
|---------------------:|-----------------------------------:|-------------:|------:|
|                PC1   |       Polycomb / Heterochromatin   |            0 |   PC  |
|                 L1   |                       Low signal   |            1 |    L  |
|                TN1   |                    Transcription   |            2 |   TN  |
|                TN2   |                    Transcription   |            3 |   TN  |
|                 L2   |                       Low signal   |            4 |    L  |
|                 E1   |                        Stem cell   |            5 |    E  |
|                 E2   |                     Multi-tissue   |            6 |    E  |
|                 E3   |               Brain / Melanocyte   |            7 |    E  |
|                 L3   |                       Low signal   |            8 |    L  |
|                 E4   |                     Multi-tissue   |            9 |    E  |
```

(continues on next page)

(continued from previous page)

```
|                 TF1  |                       NANOG / FOXA1  |     10 |    TF  |
|                 HET1 |                     Heterochromatin  |     11 |   HET  |
|                  E5  |                         B-cell-like  |     12 |     E  |
|                  E6  |                     Weak epithelial  |     13 |     E  |
|                 TF2  |                               CEBPB  |     14 |    TF  |
|                 PC2  |                       Weak Polycomb  |     15 |    PC  |
|                  E7  |                Monocyte / Macrophage  |     16 |     E  |
|                  E8  |                   Weak multi-tissue  |     17 |     E  |
|                  L4  |                          Low signal  |     18 |     L  |
|                 TF3  |                   FOXA1 / AR / ESR1  |     19 |    TF  |
|                 PC3  |                            Polycomb  |     20 |    PC  |
|                 TN3  |                       Transcription  |     21 |    TN  |
|                  L5  |                          Low signal  |     22 |     L  |
|                 HET2 |                     Heterochromatin  |     23 |   HET  |
|                  L6  |                          Low signal  |     24 |     L  |
|                   P  |                            Promoter  |     25 |     P  |
|                  E9  |                   Liver / Intestine  |     26 |     E  |
|                 CTCF |                         CTCF-Cohesin  |     27 |  CTCF  |
|                 TN4  |                       Transcription  |     28 |    TN  |
|                 HET3 |                     Heterochromatin  |     29 |   HET  |
|                 E10  |                               Brain  |     30 |     E  |
|                 TF4  |                                OTX2  |     31 |    TF  |
|                 HET4 |                     Heterochromatin  |     32 |   HET  |
|                  L7  |                          Low signal  |     33 |     L  |
|                 PC4  | Polycomb / Bivalent stem cell Enh |     34 |    PC  |
|                 HET5 |                          Centromere  |     35 |   HET  |
|                 E11  |                              T-cell  |     36 |     E  |
|                 TF5  |                                  AR  |     37 |    TF  |
|                 E12  |                   Erythroblast-like  |     38 |     E  |
|                 HET6 |                          Centromere  |     39 |    HET |
```

Regulatory feature scores

- **diffs**: The difference between the the predicted probability of the reference allele and the alternative allele for a regulatory feature ($p_{alt} - p_{ref}$).

5.7 DeepSEA (Beluga)

5.7.1 Introduction

DeepSEA is a deep learning-based algorithmic framework for predicting the chromatin effects of sequence alterations with single nucleotide sensitivity. DeepSEA can accurately predict the epigenetic state of a sequence, including transcription factors binding, DNase I sensitivities and histone marks in multiple cell types, and further utilize this capability to predict the chromatin effects of sequence variants and prioritize regulatory variants.

The 2019 version of DeepSEA, nicknamed '**Beluga**', can predict **2002** chromatin features. Beluga is described in:

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya, **Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk**. Nature Genetics (2018).

To determine if certain features (ie. transcription factors, marks, or cell types) are present/accounted for in the model, refer to the supplemental feature table which has all the profiles used to train DeepSEA.

DeepSEA is originally described in the following manuscript:

Jian Zhou, Olga G. Troyanskaya. **Predicting the Effects of Noncoding Variants with Deep learning-based Sequence Model.** Nature Methods (2015).

To determine if certain features (ie. transcription factors, marks, or cell types) are present/accounted for in the model, refer to the supplemental feature table which has all the profiles used to train DeepSEA.

5.7.2 Input

DeepSEA predicts genomic variant effects on a wide range of chromatin features at the variant position (Transcription factors binding, DNase I hypersensitive sites, and histone marks in multiple human cell types). DeepSEA can also be utilized for predicting chromatin features for any DNA sequence.

File formats

We support three types of input: vcf, fasta, bed. If you want to predict effects of noncoding variants, use vcf format input. If you want to predict chromatin feature probabilities for DNA sequences, use fasta format. If you want to specify sequences from the human reference genome (GRCh37/hg19), you can use bed format. See below for a quick introduction:

VCF format is used for specifying a genomic variant. A minimal example is `chr1 109817590 - G T` (if you want to copy cover this text as input, you will need to change spaces to tabs). The five columns are chromosome, position, name, reference allele, and alternative allele.

Fasta format input should include sequences of 2000bp length each. If a sequence is longer than 2000bp, only the center 2000bp will be used. A minimal example is

```
>TestSequence
TGGGATTACAGGCGTGAGCCACCGCGCCCGGCCCATTGTACCATTCTTAT
GCCTTTGCGTCCTCATAGCTTAGCTCCCGTATATCAGTGAGAACATACTA
TGTTTGGTTTTCCATACCCGAGTTACTTCACTTAGAATAATAGTCTCCAA
TTTCATCCAGGTCAGTGCAAATGCGTTAATTCGTTCCTTTTATGGCTGAG
TAGTATTCCATCATATATATATACTACAGTTTCTTTATCCACTCGTAAAT
TGATGGGCATTTGTGTTGGAACACTTCTCCACTGCTGGTGGGAATGTAAA
TTAGTGCAGCCACTATGGATAACAGTGTGGAGATTTGTTAAAGAACTAAA
ACTAGAACTACCATTTGATCCAGCAATCCCACTACTGGGTATCTACCCAG
AAGAAAAGAAGTCATTATTTGAAAAAGATACTTGCACGGGCATGTTTATA
GCAGCACAATTCACAATTGTAGTTGTATTTCTTTAAGCGTGTCTTTTCAA
TATCTCTCATGTTTCTGGTATAGATGGTATATATGTTAATCTTGTTCCTG
AGGTCTGTTTTTTATTTTTGTCATTAAAGTGGGAATTAAATAGTTTTGTA
GTGCATATAAATTAAAGAAAAAGTTCACATAAGCATATTTGCCAATCATC
TCAAAATGCTATATTCTCCTTCACGGTTTTGAAAATAATTCAGGGTTTTC
TCTTCCTCATTGCTTTCCCACCAACTGACAGTATTATTTTCTTAGTCATT
TTACTGACCTTTGAAATTACTCCTTTGAGGTCTTCTAAAAAATTTTATGG
GCTCTGCTGCTTTTTGGTGGCCTCCTTGTATCATTTATTCTATTACAGGA
CGACTTACAAAAGGAAGCACATAAATTGACCCATATACATATCCTATCAT
TGGGGAGTTTCTGTGCAAATGTTATTTATTGGAAGCTATTACTAAGAATT
GTAAGAAAAATAATTGGTATTGATGCAGCTAGTATGGTTCCTGTAATTAT
CGTACTCAGCCACGTAAATCATAGCTATATGTAGCCAAAGATCCATGAAC
AAAATTTCCAGTAACATCATTATAATTCAAAAGGCAGACTTTCAGAACCA
GACAGACTTGAATTTAAATTCTAGCTTTACCACACATGAATTTAACCTTG
TGGAAGGTTAACCTATCTAAACTCATGTTTCTTCATTGGTAGCTGATAAA
ATTAAGGATCATGTATATAACCACCTAGTAGAGTTGTTTAAGAAACTGTT
AGAATTCCATAAATTGTTAGTATTAATGAGTTTTTGTTGGACATGTGTTA
GGCTAGGCCACTCCTTGACCTTCATAGAGGTATGGATTATGACACAAATT
```

(continues on next page)

(continued from previous page)

```
CTAAACTGTAGGTAGGCATGGCTTTGTAGCAAGTATTAAAATAGTAAATA
TTTTATTTTTATAAGATAAATGTAAACCTTTTAAAAGTTTCATTACATTT
GTATTTATGAAATATCATCCTATATCAACTATAGAGAGAAGATCGCAAGA
AGGCAGTGGCAGCAGAGGCTCCAGTTAGGAGGCTACTAGTCCAAATACAT
TGCGATAAAAACTTGGCAAAAGGTGCTGGTAGTCTGATGAAATAAAGTAG
ATAAATTTTAGAGGTATTTATAAAATAATTAAAGAATATTCAATAATAGG
AGATATATTACCCAATAGAGTGGAGATTCAAAGATAACTCCGAAAGTTTT
TTGCTAAAGCAACATTTGGCTGTGCTATCATTTACTAAGAAAGACAACAA
GAGAGTAAAATCAAGTTTGAGGATGAAGTGAATTTATTCCTTTTTGATTG
ATACATAATTGACATGTAATAAAACCCACAATGTTAAGAGTTCGGTTTGA
TGTGCTTGACTATTTTAGGCACTGGTGTTATCACAACACAAGACAACAGA
TAGGACATTCTCAGAAAATTTTTTCATGTCCCTTTCCAGTCAGTTTCAAG
CCTTCTTTCCATGCAATAATTTTCTCACTTTGCCATTCTAGTAGGTGTGA
```

Bed format provides another way to specify sequences in human reference genome (hg19). The bed input should specify 2000bp-length regions. A minimal example is `chr1 109817091 109819090`. The three columns are chromosome, start position, and end position.

Genome coordinates

We support only `GRCh37/hg19` genome coordinates. You can use LiftOver to convert your coordinates to the correct version.

Large submissions

We recommend using the web server if you have <10,000 variants or sequences. You will experience degraded performance when submitting a larger set of sequences. In those instances, we suggest that you split the set into multiple <10,000 submissions, or run the standalone version on your local machine, or contact our group directly.

5.7.3 Output

Regulatory feature scores

- **diffs**: The difference between the the predicted probability of the reference allele and the alternative allele for a regulatory feature ($p_{alt} - p_{ref}$).
- **e-value**: E-value is defined as the expected proportion of SNPs with a larger predicted effect. We calculate an 'e-value' based on the empirical distribution of that feature's effect ($abs(p_{alt} - p_{ref})$) among gnomAD variants. For example, a feature e-value of 0.01 indicates that only 1% of gnomAD variants have a larger predicted effect.
- **z-score**: A scaled score where the feature diff score ($p_{alt} - p_{ref}$) is divided by the root mean square of the feature diff score across gnomAD variants. Note that this is "sign-preserving", i.e. a negative z-score indicates that a mutation **decreases** the probability of a regulatory feature.

Variant scores

- **Disease Impact Score (DIS)**: DIS is calculated by training a logistic regression model that prioritizes likely disease-associated mutations on the basis of the predicted transcriptional or post-transcriptional regulatory effects of these mutations (See Zhou et. al, 2019). The predicted DIS probabilities are then converted into 'DIS e-values', computed based on the empirical distributions of predicted effects for gnomAD variants. The final DIS score is:

$$-log10(DISevalue_{feature})$$

- **Mean -log e-value (MLE)**: For each predicted regulatory feature effect ($abs(p_{alt} - p_{ref})$) of a variant, we calculate a 'feature e-value' based on the empirical distribution of that feature's effects among gnomAD variants (see above Regulatory feature scores: e-value). The MLE score of a variant is

$$\sum -log10(evalue_{feature})/N$$

5.7.4 In-silico mutagenesis

Perform "In silico saturated mutagenesis" (ISM) analysis to discover informative sequence features within any sequence. Specifically, it performs computational mutation scanning to assess the effect of mutating every base of the input sequence on chromatin feature predictions. This method for context-specific sequence feature extraction takes advantage of DeepSEA's ability to utilize flanking context sequences information.

Note that ISM only accepts a sequence (FASTA file) as input.

ISM outputs effects for each of three possible substitutions of all 2000 bases, across all chromatin features.

5.8 ExPecto

5.8.1 Introduction

ExPecto is a framework for ab initio sequence-based prediction of mutation gene expression effects and disease risks. With this web interface, we provide an explorer of tissue-specific expression effect predictions. The current release contains all single nucleotide substitutions within 1kb to the representative TSS of a gene and all 1000 Genomes variants that passed a minimum predicted effect threshold (>0.3 log fold-change in any tissue).

The code for predicting expression effects for human genome variants and training new expression models is available at this github repository.

The ExPecto framework is described in the following manuscript:

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, Nature Genetics, 2018

5.8.2 Download

Predicted expression effects

This is the bulk download link of all mutation predictions.

Variation potential directionality scores

Variation potential of a gene in a tissue or cell-type can reflect the evolutionary constraint on its expression level. Specifically, we compute the variation potential directionality score as the sum of all directional mutation effects within 1kb to TSS. A negative variation potential indicates active expression and constraint toward higher expression level, and vice versa. The sum of absolute mutation effects, or the magnitudes, is predictive of tissue/condition-specificity of a gene. The variation potential directionality scores and the inferred evolution constraint probabilities can be downloaded here.

The full prediction of all 140 million mutations can be downloaded here (~125G).

5.8.3 Method Details

ExPecto uses exponential basis function-based linear models upon deep convolutional network model of chromatin effects. ExPecto predicts expression levels directly from sequence and is capable of predicting effects of sequence variations.

For detailed procedures of the prediction, the chromatin predictions were computed from DeepSEA "Beluga" per 200bp bin, and 200 bins centered at TSS (40kb region) were used as input to predict expression effects. To reduce the dimensionality for ExPecto model training, the predicted chromatin spatial patterns were summarized to spatial features by 10 exponential basis functions. The summarized spatial features and gene expression levels were used to train regularized linear models for the final step of the prediction. The representative TSSes are selected based on FANTOM CAGE data.

We also propose a path toward ab initio disease risk prediction through combining the prediction of expression effects and the estimation of evolution constraints on expression levels. For example, mutations predicted to have strong negative expression effects on a positively constrained gene are predicted to be deleterious. We estimate evolutionary constraints through systematic profiling of potential mutation effects through in silico mutagenesis. As proof-of-principle we showed that this approach can predict the disease alleles from both curated HGMD disease mutation data and disease GWASes.

5.9 SEEK - Search-Based Exploration of Expression Compendium

5.9.1 What is SEEK?

SEEK is a computational gene coexpression search engine. SEEK provides biologists with a way to navigate the massive human expression compendium that now contains thousands of expression datasets. SEEK returns a robust ranking of coexpressed genes in the biological area of interest defined by the user's query genes. In the meantime, it also prioritizes thousands of expression datasets according to the user's query of interest. The unique strengths of SEEK include its support for multi-gene query and cross-platform analysis, as well as its rich visualization features.

Cross-organism, cross platform, coexpression search

SEEK automatically prioritizes relevant datasets where patterns of coexpression are conserved across six organisms: human, mouse, worm, fly, zebrafish and yeast. Since results are simultaneously calculated for each organism, we rank each for their similarity to the query organism with regards to gene function preservation.

SEEK hubbiness correction

SEEK uses a hubbiness correction algorithm to prevent retrieving generally **hubby genes** (i.e., well connected genes, see Barabasi et al , Han et al, Xulvi-Brunet et al) that might have high coexpression to the query regardless of the query composition. For each gene in the retrieved list (such a gene is known as the target), it subtracts the average coexpression score of the target gene calculated from the coexpression of the target to all genes in the genome. The effect of this correction is that a highly connected target gene will be brought down in the ranking due to subtracting its higher average coexpression score, so that the degree of the genes will be balanced out in the coexpression network, and the search result will reflect genes that are more specifically correlated with the query.

Evaluation and example

We tested this on a group of 344 GO Biological Process slim terms, retrieving co-annotated genes from each slim term. This hubbiness correction brought improvement to 219 GO terms, with the average performance improvement being 124%.

In the other 125 GO terms where performance did not significantly improve or perform worse, the correction procedure was able to retain >83% of the original performance. The performance is measured in terms of the precision at 10% recall. In another evaluation, we sought to evaluate whether SEEK successfully downweight frequently retrieved genes.

Specifically, we checked the rank difference that the correction makes on specific genes. We searched 1000 randomly selected queries. The Table below shows the frequency that the hubby genes appear in the top 100 rank positions before and after the correction procedure.

docs/img/SEEK_hubby_genes.png

SEEK vs SPELL comparison

SPELL (Hibbs et al) is a previously developed algorithm designed to search for coexpressed genes in the yeast expression compendium. While this algorithm was helpful for yeast, it was insufficient for searching the large human compendium, which is 20-times greater than yeast (~5,000 datasets compared to 300), and the number of genes in human is also 4-times greater (25,000 compared to 6,000).

We found SEEK to be better than SPELL in terms of tackling the dramatic increase in the size of the human data, where the human genes also exhibit substantially more heterogeneous expression patterns. In SEEK, we have made many data-structure changes, optimizations to the system and implementations, using the Sleipnir library. The search algorithm is also fundamentally different from SPELL. The figure below shows that SEEK beats SPELL in 248 out of 344 evaluated GO biologial processes (when we searched a subset of each process' genes to retrieve the rest).

docs/img/SEEK_SPELL_comparison.png

The average performance improvement is 154% in precision at 10% recall. Much of the improvement comes from the cross-validated dataset weighting algorithm that is flexible to detect partial coexpression between the query genes using a robust rank-based framework. In the Figure, $n_1$is the number of GO terms where SPELL outperforms SEEK; n_2 is the count of the reverse.

5.9.2 Getting Started

Starting a search

Enter the query in gene-symbols, separated by spaces (see the Figure below). Query can be a single-gene or multiple genes (up to ~150). If the query is multi-gene, then there should be some connections between the query genes (such as coexpressions), or the query should be biologically coherent (for example, they describe a common biological process, function, module, or they physically interact).

Viewing the retrieved genes and datasets

Expression is the default view of the search results (shown below). The query genes and their coexpressed neighbor genes are displayed, and a side-by-side comparison across datasets is shown.

The top 3 datasets are automatically selected and ordered by relevance to the query genes. Above the heatmaps are the dataset titles. To the left of the heatmaps the row header are the gene names and coexpression score. The gene can be clicked to open up its HumanBase network anaylsis in a new browser tab. SEEK derives a single integrated coexpressed gene ranking, since it is more reliable than from a single dataset. This integration weights datasets differently, according to which query genes are used.

You may search the dataset titles and add or remove datasets to compare. The titles of the selected datasets appear in an expandable *accordion* component (see below), which shows the dataset details when opened.

Gene-enrichment analysis

SEEK allows users to search for a set of genes from one of six organisms: human, mouse, worm, fly, zebrafish, and yeast, to find patterns of coexpression. The SEEK system then automatically prioritizes relevant datasets, where patterns of coexpression are conserved. Since results are simultaneously calculated for each organism, we rank each for their similarity to the query organism with regards to gene function preservation. We also show term enrichment across the prioritized datasets to better understand the different experimental contexts in each model organism that are driving the observed results.

SEEK converts all genes from the initial query into their orthologs using annotations from the OrthoMCL database.

SEEK is then run for each individual organism - ranking all genes by coexpression to each query and weighting datasets where they are coexpressed. Rank-based enrichments are then calculated for the gene rankings and the datasets to give a picture of the functional similarities between organisms.

These functional enrichments for the genes are then each compared to the enrichment terms of the query organism in a pairwise manner (using Spearman correlation) that captures how many processes are shared between the query and the other organism.

Finally, these results are ranked and presented to the user along with the lists of shared GO terms derived from the gene rankings and shared terms covered by the dataset rankings (see Figure: Ortholog Ranks and Figure: Gene and dataset enrichments).

SEEK provides an avenue to explore coexpression patterns within an organism, but in addition, also allows users to examine their conservation across organisms, which can facilitate knowledge transfer between species. These cross-organism comparisons are crucial, as some particular disease processes may be more evident in the coexpression patterns of one organism versus another. In our case studies, we found that some disease processes have distinct

docs/img/SEEK_Getting_Started_1.png

Fig. 4: SEEK query component

docs/img/SEEK_Case_Study_Hedgehog.png

Fig. 5: SEEK expression view

docs/img/SEEK_Getting_Started_2.png

Fig. 6: SEEK expanded dataset panel

docs/img/SEEK-Enrichment-Flowchart.png

Fig. 7: Flow chart description of SEEK enrichment

docs/img/SEEK-Ortholog_Ranks.png

Fig. 8: SEEK ortholog ranks

docs/img/SEEK-Gene_and_Dataset_Enrichment.png

Fig. 9: SEEK gene and dataset enrichments

mappings in particular organisms, suggesting that distinct model systems can capture useful, unique facets of disease pathology.

Limit search to tissue or disease related datasets

By default, SEEK searches through the entire compendium to discover relevant datasets and coexpressed genes. However, users can limit the scope of the search to specific disease, cell, or tissue categories. This is helpful if a user wants to view expression only in a given expression context.

To limit the query this way, before you submit the query, first choose from among the tissue or disease categories listed. You will find them using the searchable **'Dataset filter'** component on the query page. Once selections from the available categories are complete, click "Submit" and SEEK will perform the query utilizing only the subset of datasets related to the chosen categories.

5.9.3 Case Studies

Case Study #1: Study a pathway of interest

This example shows how SEEK can help users to achieve these three objectives:

i. Explore a pathway across the diverse compendium datasets, in this specific example we will explore the Hedgehog signaling pathway (Hh)

ii. Find disease states and cancer types in which Hh pathway genes are coexpressed (i.e. find datasets associated with the Hh pathway)

iii. Discover other gene candidates in this pathway and examine them in the Functional Module Detection (FMD) tool which you can read about in these docs.

i. Explore a pathway

Hedgehog (Hh) pathway is a major development and cancer pathway. This pathway is perturbed in cancer patients likely caused by mutations. The pathway is SHH, DHH, IHH ligand dependent and upon ligand binding it produces the transcription factors GLI1, GLI2 which then activate a wide range of downstream processes.

To start exploring this pathway, we enter **GLI1 GLI2 PTCH1** as the query genes, which are transcription factors and receptor protein that are markers of this pathway, and central to the machinery.

The figure below shows the result of this query. In this figure, the prioritization of datasets is based on the coexpression of the query genes. The top 3 datasets are automatically selected and shown in an expandable *accordion* component, and shown as well in the 3 heatmaps arranged side by side. These prioritized datasets represent cancer studies where the expression/coexpression of the pathway genes indicate the importance of the Hh pathway activations. Expand any dataset title in the accordion to learn more about the study.

Click here to interact with this example in a new browser tab.

ii. Find disease states and cancer types

When we examine the top datasets in this example, we have simultaneously discovered Hh activations across a diverse set of disease states, such as medulloblastoma, rhabdoid tumors, lung small-cell carcinoma. Many of these have confirmed literature associations to aberrant Hh signaling [1] [2] [3] [4].

Previously, we know that Hh misregulations often result in the constitutive activation of the pathway. Here we use the coexpression of the pathway genes GLI1/2 and PTCH1 as a proxy to represent pathway activity. Coregulations of docs/img/SEEK_Case_Study_Hedgehog.png

Fig. 10: Hh query GLI1 GLI2 PTCH1. The top 3 datasets are automatically selected.

Hh genes in this case measures active pathway signaling. Retrieved datasets will show pathway expression profiles consistent with activating Hh dysfunction.

Pinpointing disease/cancer types associated with a pathway can be very useful. It can suggest a pathway-based stratification of cancer patients based on pathway profiles, which may lead to useful strategies for treating the patient by targeting the Hh pathway. By looking across thousands of datasets in SEEK, the coexpression landscape across diverse tissue/disease states can now be comprehensively examined.

iii. Discover other gene candidates in this pathway

To answer the third question, look at the row headings to the left of the heatmaps. These are the genes that are coexpressed with the query genes. These represent genes that are predicted to be associated with Hh. SEEK retrieved many currently known members of Hh machinery, such as **SMO, HHIP, BOC, and PTCH2**. One of the top ranked members that SEEK identified, KIF7 (rank 33, not displayed in the figure) is the homolog of Cos2 protein in Drosophila melanogaster, and was recently verified experimentally as a Hh regulator [5] [6].

Case Study #2: Study a differentially expressed gene-set, glean underlying pathways and processes

Investigators often wish to know what biological process and pathways are underlying a **differentially expressed gene-set** generated from an independent microarray study or RNASeq study. But for various reasons, the gene enrichment analysis sometimes msight not find any pathways, or the relevant pathways aren't detected. This could be due to factors such as heterogeneity of the gene-set, biological noises in the data, or limited number of genes to do enrichment on, etc. SEEK can offer an alternative solution by performing a **coexpression expansion** on the gene-set first.

For example, we have a set of 10 genes which represent biomarkers for the **ERBB2 subtype of breast cancer** (obtained from [7]). After trying gene-set enrichment analysis on these 10 genes, we could not obtain any significant enriched processes.

Query the following 10 genes in SEEK:

STARD3 MED24 GRB7 CEACAM6 SMARCE1 S100P FLOT2 ERBB2 TBPL1 TLK1

You can click here to explore the results in HumanBase.

SEEK returns several independent breast cancer studies as being highly ranked among thousands of studies that are databased in the compendium. This is a reassuring sign considering that this gene set is derived from breast cancer transcriptomic experiments. Investigators can check out these datasets to learn about the experimental design, selection of patient subjects, and clinical characteristics of these patients in these related studies.

Case Study #3: Find functionally related gene pairs involving the query

The metalloproteinases (**MMP2** and **MMP9**), which function together to promote cell migration and in the breakdown of the extracellular matrix, are often found in elevated expression levels in various types of cancer [9]. Investigators can use SEEK to find the substrates of these two enzymes and the proteins that these enzymes interact with.

The results of searching this query (**MMP2** and **MMP9**, click here to interact with this query in HumanBase) indicates several collagens being highly ranked (**COL1A2, COL1A1, COL5A1**), and fibronectin (**FN1**, rank 3). These findings made sense because collagens are degraded by MMPs [10], and fibronectin promotes the activation of MMPs by stimulating their secretion [11].

Other proteins that have experimental evidence of physical interactions with MMPs are also retrieved, such as thrombospondin (**THBS2** [12]: rank 38, **THBS1** [13]: rank 88), TIMP metallopeptidase inhibitor (**TIMP1** [14]: rank 16, **TIMP2** [15]: rank 61, **TIMP3** [16] : rank 60), and SERPINF1 [17] (rank 131, also known as PEDF, and is a substrate of MMP2 and MMP9). In particular, the regulation of MMPs by **SERPINF1** is important in the context of angiogenesis, and is recently described as a promising target for cancer therapy [18].

Case Study #4: Model organisms can capture different processes of cancer

One use case of SEEK is to leverage model systems to better understand human disease. In such pursuits, users might query genes that they have identified in their study, whether from a model system or clinical data.

To simulate the latter, we used SEEK to search for disease genes taken from COSMIC (the Catalogue of Somatic Mutations in Cancer). Using these we will show how mouse and fly can be used as models of pancreatic cancer.

Pancreatic cancer has one of the worst prognosis rates of any tumor type with the chance of 5 year survival at only 5%. One main contributing factor to the poor survival rate is the fact that pan-creatic cancer is often not diagnosed until it is late stage, and symptoms are non-distinct. Any clues that would enable early detection or treatment would be important breakthroughs.

We queried tier 1 human pancreatic cancer genes from COSMIC to see if we could find any interesting disease characteristics. Mouse (p=0.46) and fly (p=0.33) models are the most functionally correlated with the query. Epigenetic processes (e.g., chromatin modification, protein ubiq- uitination, and protein acetylation) are strongly enriched in both organisms, consistent with the recent studies that demonstrate the important role of epigenetic modifications in pancreatic cancer.

Both models are also enriched for datasets with ribosome descriptors (mouse p=3.3e4 , fly p=7.1e 3)). The pancreas is primarily a metabolic organ, and though fly does not have an explicit pancreas, datasets related to metabolic processes are enriched in the SEEK results for fly (glucose p=4.6e 3, type 2 diabetes p=5.4e 3, superoxide dismutase p=2.6e 3).

Mouse datasets do not have a dominating signal and are enriched for a mix of terms relating to different disorders and environmental toxins. These disorders (e.g., intrahepatic cholestasis p=5.9e 3, scleroderma p=4.7e 2) have hallmarks of pancreatic inflammation or toxins (e.g., butadienes p=1.3e-4) which have been shown to be damaging to the pancreas. These findings demonstrate that SEEK can pick up consistent signals between organisms that reflect functional features of their human counterparts.

5.9.4 Evaluating your search result

Use the Gene Enrichment function to evaluate the coexpressed genes

SEEK chooses the widely used **GO biological process** gene-sets as gold standard for the evaluation of coexpressed genes. Using the top R number of genes, users can examine enrichments in biological processes, as well as KEGG pathways, MsigDB (GSEA) gene sets. By default, SEEK will look for enrichment within the top 100 genes. However, it is possible that such enrichment may exist beyond top 100 genes (up to 500 genes). 500 genes represent approximately 2.9% (or 17K genes) of human genome ranked by SEEK, so at this depth we may get significant coexpression. Use the gene enrichment module to adjust these settings.

Note that a lack of enrichments beyond 500 genes likely means that the retrieved coexpressed genes are highly heterogeneous, possibly resulted by a heterogeneously expressed query. In this case, refinement of user's query is recommended.

The SEEK publication has done **systematic gene retrieval evaluations** for over 995 different GO biological processes.

In general, the higher the enrichment score, the better is the biolgical signal within the coexpressed genes (and so can be said about your query genes, due to the guilt-by-association principle). SEEK allows users to highlight which coexpressed genes overlapped with a given process' gene-set annotations.

Use the dataset enrichment chart to check for over-representation

Tissue or disease categories may be over-represented among top datasets prioritized by SEEK given query. Since every dataset is associated with some tissue/disease (non-cancer, cancer) terms, by checking for dataset-set enrichment, users can notice over-representations in these categories. Note that if tissue categories were selected, this is especially useful for prioritizing between tissues based on a gene-set of interest. One potential drawback is that these dataset categories may not fully capture the full spectrum of experimental conditions, as concepts such as tissue and disease terms may be general. So if users wish to be specific, it is recommended that they read the description of each prioritized dataset to fully evaluate its relevance.

A nice feature of SEEK is that it prioritizes more than 10k datasets given query genes and based on which exhibits significant coexpression. Users can check the produced list where an interesting dataset is ranked relative to the query.

Uses a clustering based evaluation

In order to assess coexpression relationships between query genes, clustering (or correlation) based measures are defined to individually evaluate datasets. If query genes are strongly clustered more so than random groups of genes in each dataset, this indicates that relevant biological processes are active and the dataset is relevant.

SEEK provides coexpression P-values for all datasets in the compendium. The measure is based on rank-biased version of Pearson correlation (see publication, referred to as the "dataset weight")

The clustering of genes offers a lot of information about the heterogeneity of query gene-set in the cancer samples. SEEK calculates, and furthermore visualizes how query genes are coexpressed with each other in the Expression Viewer. With this viewer, we can intuitively interpret large queries (ie. 10 query genes or more) where it is impossible to know what coexpressed groups may be formed within a large query.

How do I improve the results?

If you get a weak result after evaluating with the above methods, what can you do to improve your results?

1) **Refining the datasets** - perhaps you notice that the all-dataset search mode does not work very well for your query. In this case, try refining to a tissue or disease of interest.

   If you prefer a wide-reach similar to all-dataset mode but still wished to refine for instance by cancer the solution would be to refine by cancer datasets (a highly general category with over 3000 datasets). The number of datasets is listed next to each entry in the *Dataset filter*.

   If you don't know which tissue to refine to, because you don't know which tissues they are expressed in. We suggest running the query without selecting any tissues. The resulting top genes tissue your query is expressed (this works for both single gene and multi-gene query).

2) **Refining the query**

   Small query - (<3 genes). Small queries may sometimes not allow SEEK to accurately prioritize datasets. In this case, we suggest expanding your query with functionally related genes (such as those that physically interact with the query). This may improve the result. Use **STRING**, **IMP** to get these genes. Along this line, another popular approach is to add tissue or disease specific genes to your query.

   Large query - use visualization based evaluation discussed above to filter your query to a coexpressed subset.

5.9.5 FAQ

General questions

What is SEEK?

SEEK stands for Search-based Exploration of Expression Compendium. It is a gene-based human coexpression search system. Given a query gene-set, the system prioritizes thousands of expression datasets (deposited in the public repository GEO) in order to find those that may be relevant to the query. Additionally, SEEK integrates datasets to identify other genes that are coexpressed with the query genes.

What is SEEK used for?

Following are some scenarios in which finding coexpressions could be useful:

- When users define a query of a single-gene, SEEK can retrieve coexpressed genes to reveal insights about the function of the query gene.
- Biologists might have a small set of candidate genes from genetic screens, or other genomic studies. When users input them as a query gene-set, SEEK can retrieve other genes as a part of the common biological theme underlying the query gene-set (a biological process, pathway, molecular function, common miRNA or TF regulator, etc).
- The coexpressed genes may also identify possible gene-interactions involving the query.

Because SEEK prioritizes datasets, SEEK also helps to establish associations between the query gene-set and tissues, diseases, and cell-types (which are described in the dataset metadata).

You can ask questions such as:

- What are the datasets in the compendium where my query genes are coexpressed?
- Are these datasets with query coexpression seem to be associated with a particular disease or tissue type?

What are the advantages of SEEK?

Advantages include:

- Robust and cross-platform coexpressed gene integration, which means that coexpressed genes from multiple platforms can be added together to give a robust gene ranking.
- A large collection of expression datasets being used for integration (5500 datasets with 155,000 arrays, and include RNASeq datasets).
- Global or area-specific coexpression search.
- Attractive visualization of expression patterns with flexible attribute-based condition display and clustering.

Algorithm questions

What is the dataset weighting algorithm used by SEEK?

The weight of each dataset is calculated at the search time and uses the query genes. The rationale is to up-weight datasets where the query genes are coexpressed [1]. The more coexpressed they are in a dataset, the more relevance the dataset has, and the higher the weight will be.

A **cross-validation based algorithm** is being used to give robust dataset weights. This divides the query into several parts, chooses one part as a sub-query, then evaluates how well the dataset retrieves the remaining query parts.

Frequently, the query genes are only **partially coexpressed** even in the most informative datasets. As a result, the correlations between the non-coexpressed parts of the query can hurt the weight of dataset that is actually calculated from the coexpressed, informative part of the query. To solve this challenge, SEEK utilizes a rank-based procedure, inspired by **rank-biased precision** [2] from information retrieval, to give emphasis on the high correlations between genes in the query.

Since the weighting of dataset is based on the similarity of the query genes, those datasets where query genes have incoherent expression will be automatically ignored in integration (these could be low quality datasets or datasets with spurious correlations related to the query, or irrelevant datasets). Thus this algorithm achieves **automatic data quality control**.

How does SEEK compute significance for dataset weight?

The significance P-value is computed from a background distribution of random coexpression edges made from a random set of genes with the same size as a real query. Such a background distribution is specific to each dataset and to each query size. A random trials made up of 1000 random queries were used and a **generalized pareto distribution** [3] was fitted to extract parameters of the background distribution for easy computation of the P-value.

How is the score of each gene computed?

Computing the final gene score uses the dataset weights (previously discussed in this FAQ) in order to reflect the coexpressions that are located in the top relevant datasets. For each gene g, the final score is:

docs/img/SEEK_fg_formula.png

$$F(g) = \frac{\sum_{d \in D} s_d(g) w_d}{\sum_{d \in D} w_d} \quad (5.1)$$

Where D is the set of datasets that contain g. In the equation, the score of g in each dataset $s_d(g)$, is given by:

docs/img/SEEK_sd_g_formula.png

$$s_d(g) = \sum_{q \in Q} z_d(g, q)/|Q|$$

Where z_d is the correlation and Q is the query. To reduce the bias caused by those genes with insufficient dataset coverage, we discard genes that are covered by less than 50% of the compendium. These genes automatically have the lowest score.

How do I know if the coexpressed genes retrieved by SEEK are significant?

In order to assess the significance of the retrieved genes, we adopt a **null model** where we assume that the **query is random** (i.e., a random set of genes). We generated 10,000 random queries consisting of 100 queries per query-size, where size ranged from 1 to 100 genes. We searched all random queries in SEEK and produced a set of gene-rankings. Given a true query, to estimate the significance of gene `x` in the true query's ranking, we estimate the fraction of random queries where the rank of `x` is higher than the rank of `x` in the true query. We note that the null model is generally very similar between different query sizes beyond the query size of 10 genes. So we can use a size-free estimation for these query sizes.

How do I know if my query is coexpressed or not?

Since the dataset weight is calculated by query coexpression, the dataset weight can directly answer this question. In general, the query would be considered coexpressed if there is a subset of datasets in the compendium with sufficiently high dataset weight.

The **significance** of the dataset weight can indicate how query coexpression is compared to random. The **number** of datasets with significant dataset weight (given some P-value threshold) can indicate whether this query coexpression is widely occurring in the compendium or restricted to a subset of datasets.

What is a dataset keyword?

A **keyword** is a curated term (in a controlled vocabulary) that describes a dataset. In SEEK, keywords come from the UMLS controlled vocabulary , which specifies a comprehensive set of tissue, disease types. To determine what keywords are annotated to each dataset, SEEK uses a semi-automatic strategy that involves text-mining followed by manual curation. The text-mining mines for controlled vocabulary terms within dataset description and sample description texts associated with the dataset. In manual curation, we review and correct the mappings for those commonly mismapped keywords.

Usage questions

How do I narrow down the scope of datasets used in the query?

SEEK by default utilizes ALL of the thousands of datasets in the compendium for the query search. Users can however restrict to datasets with particular characteristics, such as disease-type, tissue-type, etc.

To focus your query use the **Dataset filter** on the **Query** page. For example, to restrict the query to datasets with keyword ‘Brain’, you can type ‘Brain’ in the **Dataset filter** box and a list of biological terms filtered by that keyword will be shown. You can then select as many terms of interest as you like. Only datasets from those terms will be considered when you submit the query.

docs/img/SEEK_Getting_Started_3.png

How do I get the complete list of genes or datasets prioritized to the given query?

On the SEEK expression result page, next to the heatmap legend there is a button labeled *Download*. Clicking on this button will allow you to choose between downloading a CSV of either the genes ranked by coexpression score or datasets ranked by query relevance (aka weight).

How can I check the rank for a gene or dataset of interest?

There are two ways to check the rank for a gene or dataset of interest:

1) Get the complete list of coexpressed genes or datasets (see previous question) and search for your gene / dataset of interest in the CSV. The rank is included in the first column of each row.
2) (Gene only) On the SEEK expression result page, there is an option panel with the label *Filter expression results by gene*. You can paste a list of genes which you are interested in and HumanBase will filter the list of genes displayed in the heatmap to only those genes of interest. The rank is included on each row of the filtered heatmap.

How can I visualize the expression for a particular gene of interest?

See #2 in the previous question.

Which datasets were used for my query?

SEEK by default considers all of the thousands of datasets in the compendium for the query search (approximately 10,600). Datasets are weighted according to which query genes are used. To review the list of datasets used in a specific query, on the *Co-expression results* tab either:

1) Click on the *Download* -> *Download datasets CSV* button to download a CSV of the datasets used in the query.
2) In the *Selected datasets shown* section, instead of typing title text, click on the down arrow to the right of the search box. This will open a list of all of the datasets used in the query. You can then select any datasets of interest and they will be added to the datasets in the heatmap.

How large a query can SEEK handle?

SEEK can accept both single-gene and multi-gene queries. While queries involving several hundreds of genes are technically feasible, we do not recommend using such large queries, because they are likely to have heterogeneous expression patterns, which can contribute to a poor result. Such queries also consume large amounts of resource and take longer to complete. We therefore recommend queries with 150 genes or less.

How much time does searching a query take?

The time depends on the size of the query and the volume of traffic. If the server is not busy, the search speed is approximately 3 seconds per query gene and the time scales up linearly for larger queries. For example, searching a 3-gene query takes about 9 seconds.

5.9.6 Citation

Targeted exploration and analysis of large cross-platform human transcriptomic compendia Qian Zhu, Aaron K Wong, Arjun Krishnan, Miriam R Aure, Alicja Tadych, Ran Zhang, David C Corney, Casey S Greene, Lars A Bongo, Vessela N Kristensen, Moses Charikar, Kai Li & Olga G Troyanskaya Nature Methods (2015) (paper link | PMID: 25581801)

5.10 Citations

5.10.1 Tissue-specific networks, NetWAS

Greene CS*, Krishnan A*, Wong AK*, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T, Troyanskaya OG. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics. 10.1038/ng.3259w.

5.10.2 Variant effect predictions (ExPecto)

Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, and Troyanskaya OG. (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, Nature Genetics.

5.10.3 Autism gene predictions

Krishnan A*, Zhang R*, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG.(2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature Neuroscience.

5.10.4 Tissue-expression predictions

W. Ju#, C.S. Greene#, F. Eichinger, V. Nair, J.B. Hodgin, M. Bitzer, Y. Lee, Q. Zhu, M. Kehata, M. Li, M.P. Rastaldi, C.D. Cohen, O.G. Troyanskaya*, and M. Kretzler*. Defining cell type specificity at the transcriptional level in human disease. Genome Research. 23:1862-1873. 2013